

Linguistics and Literature Review (LLR)

Volume , Issue 2, October 2015

Journal DOI:

Issue DOI:

ISSN: 2221-6510 (Print) 2409-109X (Online) Journal homepage: <http://journals.umt.edu.pk/llr/Home.aspx>

Towards Transliteration between Sindhi Scripts Using Roman Script

Mehwish Leghari
Mutee U Rahman

To cite to this article: Mehwish Leghari & Mutee U Rahman (2015). Towards Transliteration between Sindhi Scripts Using Roman Script, *Linguistics and Literature Review* 1(2): 95- 104.

To link to this article:

Published online: October 31, 2015

Article QR Code:



A publication of the
Department of English Language and Literature
School of Social Sciences and Humanities
University of Management and Technology
Lahore, Pakistan

Towards Transliteration between Sindhi Scripts Using Roman Script

Mehwish Leghari

Mutee U Rahman

Department of Computer Science, Isra University - Hyderabad, Pakistan

ABSTRACT

In this research a model for transliteration is presented for two scripts of Sindhi language that is Perso-Arabic script and Devanagari script, based on an intermediate Roman script. After analyzing both Perso-Arabic and Devanagari scripts, a set of Roman script for Sindhi language is also suggested. Different issues, complexities and problems of Sindhi transliteration are discussed in detail. An algorithm to transliterate between two scripts of Sindhi language is also proposed.

Keywords: Perso-Arabic script, transliteration, complexities, Devanagari script, Roman script

Introduction

Transliteration is transformation of text from one script to another, usually based on phonetic equivalencies (IBM, 1999). Popularity and simplicity of roman script is a major reason behind the motivation of transliteration of language scripts. People and software are getting benefited from these transliteration aids. Transliteration from native scripts to Roman script has been achieved for many Asian languages including Arabic, Bengali, Persian, Hindi, Punjabi and Urdu. Transliteration between Punjabi scripts (Malik, 2006 b) and Hindi to Urdu transliteration (Malik, et al., 2008) are key examples of South Asian language transliterations. Different transliteration applications like Google Transliteration IME facilitate the users of different languages to transliterate from their native scripts to Roman script. It is currently available for 19 different languages including Arabic, Bengali, Farsi (Persian), Gujarati, Hindi, Punjabi, Sanskrit and Urdu. Transliteration of two scripts of Sindhi has not been achieved yet and in fact not even initiated. Sindhi computational linguistics has not received much encouragement in either Pakistan or India (Khumbhandani, 1970) and works are limited to font design and word processing. But Sindhi computing should not only encircle font design and word processing but extensive research is

CONTACT Mehwish Leghari at legharimehwish@hotmail.com

needed in the areas of artificial intelligence, computational linguistics, natural language processing, corpus linguistics and script processing (including transliteration) (Rahman, 2009).

Following sections discuss a brief history of Sindhi language, Perso-Arabic and Devanagari scripts of Sindhi language, composition of both scripts, a set of suggested Roman script for Sindhi language and an algorithm for transliteration between two scripts of Sindhi language.

Sindhi language

Sindhi is an Indo-Aryan language with its roots in ancient history. Sindhi is being spoken by approximately 40 million (Sindhi Language Authority, 2009) people in Sindh province of Pakistan as well as in several states of India. In Pakistan Sindhi is written in Perso-Arabic script while in India Sindhi is written in both Devanagari and Perso-Arabic scripts. Both regions heavily share the same vocabulary.

Sindhi scripts

Sindhi scripts and their writing systems are briefly described below.

Perso-Arabic script

The Perso-Arabic script of Sindhi language consists of 52 letters; most of those are taken mainly from Arabic alphabet, some letters from Persian and few modified letters. In Perso-Arabic script each letter has one to four forms according to its position beginning, middle, final and standalone. Letters in Perso-Arabic script are divided in different types on the basis of phonemes. These different types are discussed below with reference to their phonemes, writing style and position in a word.

First type is aspirated consonants. In Perso-Arabic script some of the aspirated sounds are written by combining two letters. For example aspirated form of گ can be written by combining گ (g) with ه (h) as گھ (gh) similarly there are some other aspirated consonants. In Roman script “h” is combined with the letter of nearby sound. For example “g” + “h” = “gh” is used to represent گھ (gh). On the other hand there are some aspirated consonants in Sindhi that are represented using a single letter like “چ” (chh) and “ٺ” (th).

The non-aspirated consonants of Perso-Arabic script are transliterated according to their phonemes. At some places we find multiple non aspirated consonants for single phoneme and these all have single counterparts in Devanagari script. These are further discussed in section 3.

There are three main vowels in Perso-Arabic script those are ا, و, and ي these vowels when come at the beginning of a word are simply treated as non-aspirated consonants; and at the end of the word these are treated as vowels. But in the middle of a word they need to be tackled in context of nearby letters whether those are vowels or consonants.

The diacritical marks are essential for correct accent and if missing not only creates ambiguity in transliteration but also cause misinterpretation of the words. Thus diacritical marks are equally important for avoiding ambiguities in transliteration as those are important for natural language processing and speech synthesis (Malik, et al., 2009). Some of the examples of diacritical marks

include Zabbar, Zer and Pesh. Examples of pairs of words that differ in meaning only because of difference in diacritical marks are shown in table 1.

Table 1. Effects of diacritical marks on meaning.

Word	Meaning	Word	Meaning
اَٹ	Eight	اٹ	Camel
چَپ	Lip	چپ	Silent
کَل	Laughter	کَل	Skin
مَلَن	To rub	مَلَن	To meet
کَرَن	To do	کَرَن	To fall

In Sindhi, there are four implosive stops. Using Perso-Arabic script those are represented by adding extra “dot(s)” to the letter of nearby matching sound. For example implosive version of گ (g) is گَ (gg) similarly there are three more implosive stops or implosive sounds. In Roman script these letters are represented by doubling the letters of nearby sounds like: “bb”, “jj”, “dd”, “gg” as this convention is commonly used in Sindhi-Roman script.

Devanagari script

Devanagari script adopted from Sanskrit system of writing in which each character represents a syllable. Devanagari script is written from left to right. Many of the letters having same phonemes or sounds in Perso-Arabic script are equivalent of a single letter in Devanagari script. It is also worth mention here that two letters of Perso-Arabic script have no equivalent in Devanagari script those are ء (‘a) and ع (A).

Devanagari script also have aspirated and non-aspirated consonants but unlike Perso-Arabic script the Devanagari script do not use composite letters for aspirated consonants. Hence all aspirated consonants are denoted by a single letter. In Devanagari script two types of vowels are independent and dependent vowels. Independent form of a vowel is used at the beginning of a word and dependent form is used at the end of a word. While in the middle, usually dependent form is used but there are some exceptions.

Unlike Perso-Arabic script the diacritical marks are not optional in Devanagari script. As shown in Example 1.

Example 1

तू सुहिणी आहीं.

tooN suhiNree AaheeN.

You beautiful are

تُون سُہیٹی آہین.

You are beautiful.

As shown in table 2 diacritical marks are most widely used and are integral part of words written in Devanagari script. Thus transliteration accuracy is assured while going from Devanagari to Perso-Arabic script.

Table 2. Diacritical marks in Devanagari

Devanagari	Roman	Perso-Arabic
त+ू+ं = तूँ	toon	تُون
स+ु+ह+ि+ ण+ी = सुहिणी	suhiNree	سُوْهِی
आ+ह+ी+ं = आहीं	AaheeN	آهین

Table 3. Simple consonants and independent vowels in Devanagari, Roman and Perso-Arabic scripts.

Devanagari	Roman	Perso-Arabic
आ	Aa	آ
अ	a	ا
ब	b	ب
भ	bh	پ
थ	th	ت
ट	T	ٲ
ठ	Thh	ٲ
प	p	پ
ज	j	ج
झ	jh	جھ
ञ	nn	چ
च	ch	چ
छ	chh	چھ
ख	khh	خ
द	d	د
ध	dh	ذ
ड	D	د
ढ	Dh	دھ
र	r	ر
ड़	R	ر
श	sh	ش
ग	G	غ
फ़	f	ف
फ	ph	ف

क़	q	ق
क	k	ک
ख	kh	ک
ग	g	گ
घ	gh	گھ
ङ	Ng	گ
ल	l	ل
म	m	م
न	n	ن
ं	N	ن
ण	Nr	ٹ
व	v	و
य	y	ي

Perso-Arabic, Devanagari and Roman scripts

Roman script is based on the alphabet developed by the ancient Romans, and used by most of the languages of Europe, including English, French, and German (SIL International, 2003). To achieve Sindhi transliteration an intermediate Roman script is used. As writing Sindhi and other languages in Roman script (English) is very common nowadays, so in this model, all possible steps are taken to preserve most common Roman style of writing. Therefore one do not feel any difficulty in adopting this method and can transliterate in any direction from Sindhi to Sindhi (Perso-Arabic, Roman and Devanagari or vice versa). Most of the consonants are transliterated to their matching sounds in Roman script.

Perso-Arabic, Devanagari and equivalent Roman script mapping is shown in table 3. Table 3 contains all consonants except those having same phonemes with others and implosive stops. Four unique implosive stops in Sindhi are shown in table 4.

Table 4. Implosive stops

Devanagari	Roman	Perso-Arabic
ब़	bb	ب
ज़	jj	ج
ड़	dd	د
ग़	gg	گ

Besides the letters listed in table 3 and table 4 there are multiple letters in Sindhi, using Perso-Arabic script with same equivalent in Devanagari script, as shown in table 5. This is because of similar sounds, though we have suggested separate equivalents in Roman script for these letters.

The letter ع came in Sindhi alphabet from Arabic. Native Sindhi speakers do not pronounce it properly (generating sound from inner throat) in normal conversations and there is no equivalent of ع (A) in Devanagari script (Malik, 2006a). Same is true for ء ('a) of the Perso-Arabic script.

These two letters can be transliterated easily from Perso-Arabic to Roman script and vice versa. In case of Devanagari transliteration these are either ignored or transliterated into अ (a) or े (e). Mostly these letters are ignored during transliteration of Perso-Arabic or Roman script to Devanagari script as shown in example 2 and table 6.

Example 2.

معاف کجو!

mAaaf kajo!

forgive do

माफ़ कजो !

Do forgive! / Forgive me!

Table 5. Perso-Arabic multiple consonants with same Devanagari equivalents

Devanagari	Roman	Perso-Arabic
त	t	ت
त	Tt	ط
ह	H	ح
ह	h	ه
ज़	Z	ذ
ज़	z	ز
ज़	zz	ض
ज़	Zz	ظ
स	s	س
स	S	ص
स	c	ث

Table 6. Letters with no equivalents

Devanagari	Roman	Perso-Arabic
म	m	م
-	A	ع
ा	aa	ا
फ़	f	ف

We can clearly illustrate from example 2, by further analyzing the word معاف (mAaaf) in table 6 that the use of ع (A) has been completely omitted in transliteration from Roman / Perso-Arabic to Devanagari. Those two letters are separately shown in table 7, with their equivalent Roman letters.

Table 7. Perso-Arabic letters with no Devanagari equivalent

Devanagari	Roman	Perso-Arabic
------------	-------	--------------

-	A	ع
-	‘a	ء

There are two special words that are written in some special form by using single letter and two elongated quotation marks beneath the letter. These are shown in table 8 in all three scripts. Note that roman representation of these letters is capitalized to avoid ambiguity with other letters in the word or sentence.

Table 8. Special single letter words of Sindhi

Devanagari	Roman	Perso-Arabic
ऐ	AEN	اين
में	MEN	مين

The dependent vowels and diacritical marks are shown in table 9 in all three scripts. These are shown in combination with letter “ज” (j) in Devanagari, “j” in Roman script and “ج” (j) in Perso-Arabic script to make it more understandable.

Table 9. Dependent vowels and diacritical marks

Devanagari	Roman	Perso-Arabic
ज	ja	ج
ज + ि = जि	ji	جِ
ज + ु = जु	ju	جُ
ज + ो = जो	jo	جو
ज + ू = जू	joo	جُو
ज + े = जे	je	جي
ज + ी = जी	jee	جِي
ज + ा = जा	jaa	جا

Sample conversions and problems

The dictionary lookup is used for transliteration of the words, in which one or more letters follow none of the rules. The conversions from one script of Sindhi to any other script can be achieved by implementing the rules given below:

- Take a whole word as input.
- Dictionary lookup for especial words.
- If not in dictionary, start transliterating.
- Transliterate the first letter as a consonant or independent vowel.
- From second to second last letter if any letter is consonant or vowel with a diacritical mark, transliterate it as consonant.
- If the letter is a vowel and it has no diacritical mark, transliterate it as dependent vowel.
- If last letter is vowel, transliterate it as dependent vowel.

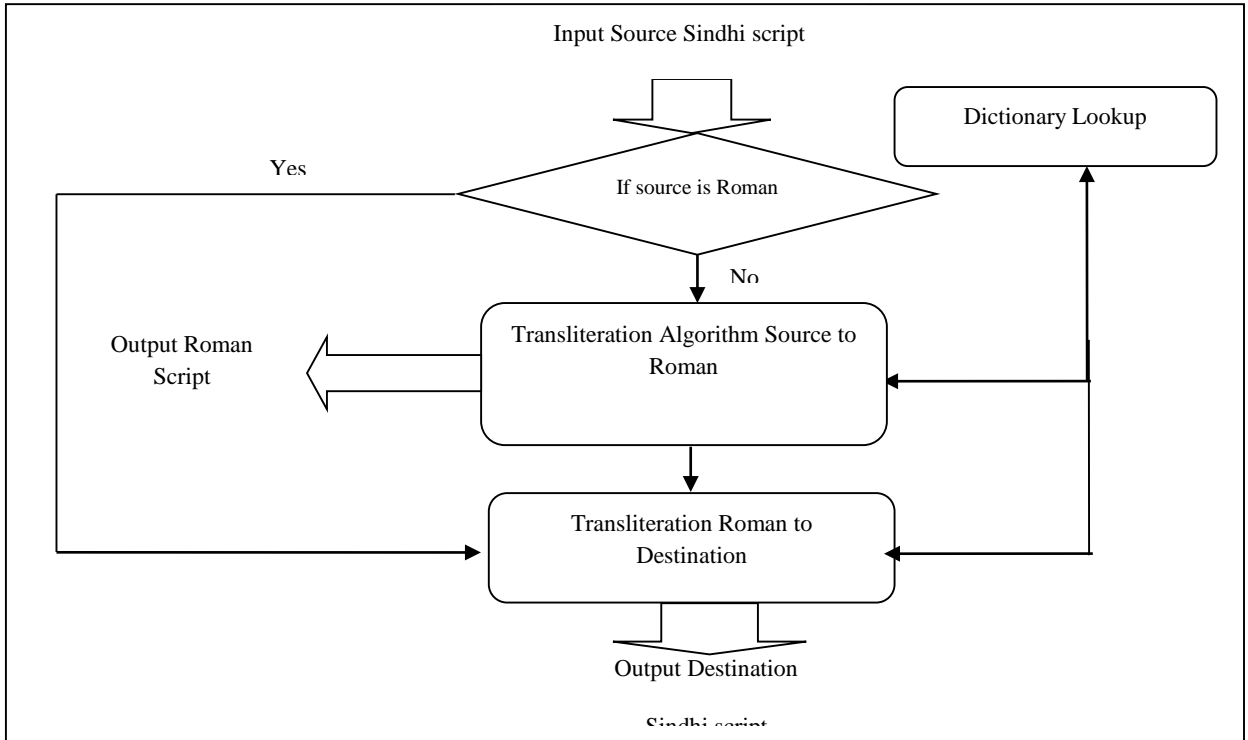


Figure 1. Transliteration model for Sindhi scripts.

These rules get more complex while transliterating from Perso-Arabic to Devanagari when there are words without proper diacritical marks. In this situation transliteration is done by analyzing the letters that come before and after the letter that have no diacritical marks. If there is no match with any condition (for example in case of consecutive vowels) then finally transliteration can be achieved on probability basis. A sample transliteration is shown in example 3.

Example 3.

बाहिर डाढी गरमी आहे.

bbaahir ddaaDhee garmee Aahe.

outside very hot is

ٻاهر ڏاڍي گرمي آهي.

It is very hot outside.

The suggested set of rules transliterates majority of sentences correctly like in example 3. However there are some ambiguities for the letters that are not properly present in Devanagari script. For Example the letter ء ('a) of Perso-Arabic script is sometimes equivalent of अ (a) in Devanagari, while अ (a) is actually equivalent of ا (a) of Perso-Arabic script. Similarly same ء ('a)

is sometimes equivalent to ے (e) (equivalent for Perso-Arabic ي (e)) and sometimes the ء ('a) is completely omitted to achieve the correct transliteration. As we can see in example 4 the letter ا (a) of first word is wrongly transliterated into letter ل (a) of the Perso-Arabic Script while its correct transliteration would be ('a) ء.

Example 4.

हूअ डाढी पियारी आहे.

hoo'a ddaaDhee piyaaree aahe.

she very lovely is

هُوا دَادِي پِيَارِي آهي. (incorrect)

هُوءَ دَادِي پِيَارِي آهي. (correct)

She is very lovely.

Conclusion and future work

After successful implementation of transliteration model discussed, one would be able to transliterate Sindhi from one script to another. People familiar with one script, will be able to understand the writings in other script. It will also be useful in implementing simple Roman to Sindhi (any of two scripts) transliteration. By implementation of successful transliteration design, the transliteration aids like Google Transliteration IME would be able to use Roman to Sindhi mapping, to make it possible to transliterate between Roman and native Sindhi scripts. Automatic transliteration will help to end up the discussions and dispute of Roman script adoption for Sindhi language. The proposed model needs to be checked on large scale by applying the algorithm on a reasonably large corpus.

References

- IBM. 1999. Glossary of Unicode terms, 2010. Retrieved from <http://www.ibm.com/developerworks/library/glossaries/unicode.html>
- Khubchandani, L. M. 1970. Sindhi In *Current Trends in Linguistics* S. T. Albert 1(1): 219. The Hauge, Natherland: Mutton & Co.
- Malik, M. G. A. 2006. *Hindi Urdu Machine Transliteration System*, (Unpublished MS Thesis, Department of Linguistics, University of Paris 7, Denis Diderot, 2 Place Jussieu, Paris France).
- Malik, M. G. A. 2006. *Punjabi Machine Transliteration*, In the proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Manchester, UK. Retrieved from <http://acl.ldc.upenn.edu/P/P06/P06-1143.pdf>
- Malik, M. G. A., Boitet, C., and Bhattacharyya, P. 2009. *A Hybrid Model for Urdu Hindi*

- Transliteration* In Proceedings of the *2009 Named Entities Workshop, ACL-IJCNLP 2009*, 177–185. Retrieved from <http://www.aclweb.org/anthology/W/W09/W09-3536.pdf>
- Malik, M. G. A, Boitet, C, and Bhattacharyya, P. 2008. *Hindi Urdu Machine Transliteration using Finite-state Transducers*, proceedings of COLING, Manchester, UK. Retrieved from <http://www.aclweb.org/anthology-new/C/C08/C08-1068.pdf>
- Rahman. M. 2009. Computational Linguistics and Sindhi Language *Sindhi Boli* Sindhi Language Authority, Hyderabad.
- SIL International. 2003. NRSI: Computers & Writing Systems. 2010. Retrieved from http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&cat_id=Glossary
- Sindhi Language Authority. 2009. *Sindhi Language 2010*. Retrieved from <http://www.sindhila.org/Sindhi Language.htm>